Master's Thesis

MSc Energy for Smart Cities

Assessment of energy efficiency savings in tertiary buildings using statistical learning techniques

REPORT

Author: Benedetto Grillone
Advisor: Prof. Andreas Sumper
Co-Advisor: Dr. Stoyan Danov (CIMNE)

Session: June 2018

ETSEIB

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

UPC

Abstract

It is estimated that about 40% of worldwide energy use occurs in buildings [1]. Increasing energy efficiency in the building sector has become a priority worldwide and especially in the European Union. It is clear that an immense energy efficiency potential lies in buildings and it is not properly harnessed. The energy efficiency increase can be realized through energy retrofitting actions, optimization of the building control strategy, or through the timely reporting of abnormal energy performance. In this thesis, a framework for the evaluation of the impact of energy retrofitting measures, with a statistical learning approach, is proposed. The model was developed as part of EDI-Net, a Horizon 2020 project, with the main goal of facilitating energy consumption monitoring in buildings and allowing analysis and evaluation of applied energy efficiency measures (EEM). The baseline models for the impact evaluation are generated using Generalized Additive Models (GAM), enhanced with auto regressive terms. Three different pilot buildings (one in Spain and two in the UK) are examined and their savings evaluated through the analysis of hourly smart meter consumption data and weather data. The results show that it's possible to evaluate energy savings in tertiary buildings using a data-driven approach, although further work is needed, in order to validate and automatize the model.

Contents

Glossary

- ANN: Artificial Neural Network
- API: Application Programming Interface
- BEMS: Building Energy Management System
- CDD: Cooling Degree Days
- DB: Data Base
- EPB: Energy Performance of Buildings
- ESCO: Energy Service Company
- GAM: Generalized Additive Model
- HDD: Heating Degree Days
- ICT: Information and Communication Technologies
- IoT: Internet of Things
- IPMVP: International Performance Measurement and Verification Protocol
- REST: Representational State Transfer
- RSS: Residual Sum of Squares
- TSS: Total Sum of Squares

Chapter 1

Introduction

Poor energy performance of the building stock is one of the main challenges to the successful implementation of energy renovation and energy efficiency strategies in Europe [2, 3]. This is one of the critical issues to be addressed before rolling out a massive strategy to reduce the energy consumption in buildings. Poor building performance is related to the building design and construction, the building materials, the mechanical and electrical systems and the control and operation of the buildings. In the case of commercial and public buildings, the application of energy efficiency measures (EEM) and retrofitting actions has a substantial impact, but no standardized method has been adopted yet, to evaluate this impact. A wide range of technologies is now available to improve the energy performance of existing buildings, but it is still a major challenge to identify the most effective retrofit measures, according to the building characteristic, as Ma et al. pointed out [4].

For this reason, this research proposes a data-driven approach that makes use of big data analytics to evaluate energy retrofit impact on tertiary buildings. The method stands in the framework of the International Performance Measurement and Verification Protocol (IPMVP) and makes use of smart meters and weather stations data.

1.1 Scope of the project

This thesis aims at developing a method that makes use of advanced statistical models, such as Generalized Additive Models (GAM)) [5] , that are able to process hourly and subhourly consumption data and evaluate their dependence on different exogenous variables, with the goal of assessing the impact of applied EEMs. The successful implementation of such a method would open several interesting possibilities, as it would allow an easy and low-cost evaluation of EEM impact, without the need of any simulation software or energy audits. Furthermore, its development in a big data environment integrates perfectly with the subsequent recommendation generation process, through which, based on building characteristics and applied measures' impacts, it is possible to detect which measures would have the highest impact on given buildings.

To the best of our knowledge, this is an innovative approach, made possible by the combination of the streaming capacities of a non-relational data base (MongoDB) with the high

storage capacity of a distributed big data data-base (Hadoop Distributed File System). For the last years, the rapid penetration of smart meters and the growing digitization of the energy systems has generated an explosion of available data which entails an increased complexity in the data processing to obtain valuable information [6]. The transformation from specifically tailored energy monitoring projects to open data projects using multiple IoT devices, is ongoing. To ensure the transition, a fundamental role is played by sophisticated ICT architectures, capable of combining real-time and batch processing data flows and of allowing a seamless connection between visual frameworks and data driven analytic tools [7].

1.2 Objectives of the project

Here, the list of objectives for this thesis is presented:

- develop a module to automatically collect and organize in a single data-frame all the relevant information for the analysis,
- create and implement, in R programming language, a logical framework, that allows to assess which are whether a measure is eligible for evaluation or not,
- implement a statistical model able to generate a baseline for the considered building, and therefore evaluate the effect of applied energy efficiency measures,
- making use of the statistical model results, calculate total impact and savings for a given EEM.

The research methodology is presented in Chapter 3, while a detailed discussion of how the different objectives were carried out can be found in Chapters 4 to 6

Chapter 2

EDI-Net structure overview and specific objectives

In this chapter, the structure of the EDI-Net application will be briefly described, since a good understanding of its functioning will enable the reader to better follow the matters described in the following chapters.

2.1 EDI-Net project and partners

EDI-Net (Energy Data Innovation Network) is a European project that received funding in the framework of the Horizon 2020 Programme. The project started in 2016 as a collaboration between seven different partners:

- De Montfort University (UK)
- Leicester City Council (UK)
- Climate Alliance (DE)
- Stadt Nurnberg (DE)
- Empirica (DE)
- CIMNE (ES
- Generalitat de Catalunya (ES)

The project is in its third and last year and the partners are now working on the extension of the services to external institutions. Before describing the application features more in detail, some numbers are presented to provide an overview of the size and kind of data that the platform is managing.

Figure 2.1: The EDI-Net logo

Table 1 shows the number of buildings currently present in the platform, the number of buildings with hourly data and the number of Energy Efficiency Measures (EEM) uploaded in the system, while in Table 2 and Table 3 we can respectively see the buildings and the energy efficiency measures sorted by category.

Entity Name	Buildings	Hourly Data Buildings	Energy Efficiency Measures
Generalitat de Catalunya	1173	136	1912
Valencia Institute of Building	303	0	5
Leicester City Council	207	201	5
Nurnberg	36	34	52
De Montfort University	24	23	0
Total	1743	394	1974

Table 2.1: EDI-Net buildings details

Cultural	Education	Healthcare	Hotels and Restaurants
93	181	546	23
Office	Residential	Sports	Others
563	147	18	397

Table 2.2: EDI-Net Buildings by Category

Heating	Cooling	Lighting	Domestic Hot Water
540	440	447	142
Electrical Equipment	Management	Envelope	Total
106	120	179	1974

Table 2.3: Energy Efficiency Measures by Category

some comment about the numbers shown in the tables.

2.2 EDI-Net system architecture

In this section, the technical details of the project will be briefly introduced. The EDI-Net application is intended to be a repository where to store the energy efficiency measures implemented in different buildings of an organization. Its objective is to facilitate the monitoring and evaluation of the implemented measures and encourage the exchange of experiences among the different users. Below, the four main functional requirements of the application are presented:

- track energy performance in detail,
- communicate energy performance in a user-friendly manner,
- facilitate comunication betweeen stakeholders,
- manage an intervention plan for energy efficiency.

The core of EDI-Net is the analysis of smart meter data from buildings, from renewable energy systems and from building energy management systems (BEMS), using Big Data analytics technologies.

The EDI-Net system has an architecture based on two big interfaces. The first interface is defined as the **EDI-Net App** and contains the user and the communication interfaces. The second interface is the **Big Data Engine** which contains the distributed storage data base and the data analytics software modules.

The main function of the EDI-Net APP is to allow the interaction between the end-users (building managers) and the Big Data Engine. It facilitates the settings editing, the data importing and the visualization of results. It is programmed under the *Django* framework. Django is a collection of Python libs allowing users to quickly and efficiently create a quality Web application. The EDI-Net APP is divided in two levels: the *Front End* and the *Back End*.

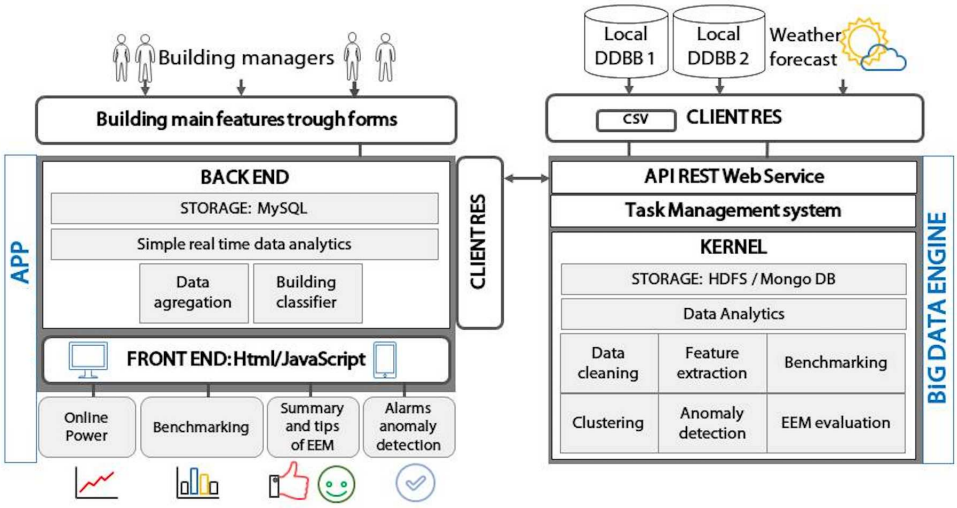

Figure 2.2: Architecture of the EDI-NET big data platform

Back End. The back end comes with an Object Relational Mapping (ORM) that allows to manipulate data sources with ease. It provides an abstraction layer, the models for structuring and manipulating the data and facilitates the creation of forms to process user input and validate data and signals. The back end is made up of two components: The *storage* and the real time *data analytics*. The storage is a structured open source data base,based on MySQL. It is linked to the REST client communication layer and stores the information introduced by the building managers through the web forms and the treated information coming from the big data engine.The real time data analytics is formed by several Django query sets which execute data manipulations on the database. Data aggregation, in several time frequencies, as well as a building classifier based on the categories defined by the building managers, constitute the main data operations performed by the EDI-Net App.

Front End.The front end helps with data selection, formatting and display. It features URL management, a language template, authentication mechanism, cache hooks and various navigation tools such as paginators. The front end support JavaScript (Js) programming language and shows, interactively, the data streaming and the results of the Data Analytics modules.

The Big Data Engine is designed to tackle the following IT challenges:

- to offer a high degree of quality of the delivered services,
- to provide batch-processing data analytic services,
- to ensure data privacy and security.

It is based on the IT architecture developed within EU funded project *Empowering* [8]. The Big Data Engine is a Representational State Transfer (REST) framework entirely developed by using open source software. It is divided into 3 levels: API REST WebService, Task Management Service and Kernel.

API REST WebService. This is the communication interface between the REST Client of the EDI-Net App and the other local databases which provide the smart meter readings of the buildings. The API is fully developed following the REST standard. The aim is to enable a Service Oriented Architecture (SOA), offering specialised energy services to the building managers. The main functions of the API are to import data into the Engine and to export data from the Engine. These objectives are addressed using different technologies. Data import and export are enabled using the *Eve* framework to implement the Web service. OpenAM provides open source Authentication, Authorization, Entitlement and Federation software. The Flask and Python modules implement all the server functionalities in order to deploy a web API server.

Task Management Service. This level is in charge of scheduling and synchronizing the tasks in the engine by means of RabbitMQ and Celery. The scheduler picks up the new task to be executed into the Engine according to a scheduling policy. The Quality of Service (QoS) is based on task return time and not on response time. Therefore, when a task enters the engine, it continues its execution in a batch mode until finalisation. Celery is the scheduler itself. RabbitMQ is a fast internal message-queuing system used to interchange information between tasks with different paradigm technologies.

Kernel. The Kernel is made up of two main components: *Data Analytics* and *Storage*. Data Analytics is a comprehensive set of modules that enable the simultaneous parallel processing of big quantities of data in order to generate the required results in a reasonable time, in other words, it is the set of modules in charge of processing the stored data. It is a combination of R and Python software libraries to allow complex calculations and data-mining tools in a Big-data environment. The combination of both packages permits the calculations to be optimized and the data processing time, reduced. Storage is the system that allows the large amount of data produced to be stored and managed in a scalable way. It is made up of a combination of low-cost hardware and database technologies that allows the acquisition, allocation and extraction of big quantities of data in a scalable manner to be processed by the analytics module. The short-term DB uses MongoDB and the long-term DB uses a Hadoop storage technology. The short-term database is used to buffer storage for data reception and sending in fast environments. It is the first data storage directly connected with the API and provides high communication bandwidth and scalability performance. It supplies temporary storage, acting as a cache memory, prior to it being permanently stored in the long-term database. Hadoop is the long-term Big-data storage where all the historical data is stored and the analytical modules operate.The main benefit of Hadoop is that it allows the process and analysis of large volumes of unstructured and semi-structured data in a cost- and time-effective way.

2.3 EDI-Net application overview

Following, the different features of the application will be briefly presented with some screen-shots of the platform, to allow easier understanding of the following chapters of the thesis.
The general dashboard of EDI-Net allows users to access a list of all the buildings they are managing. From this page, it is possible to order the list of buildings according to the consumption, the efficiency, or the number of energy efficiency measures applied.

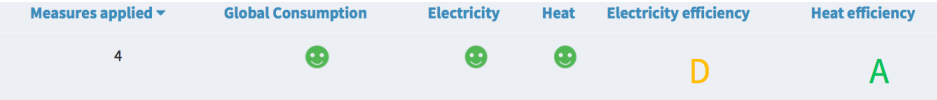

Figure 2.3: The EDI-Net dashboard

The smileys shown in the 'global consumption', 'electricity' and 'heat' categories are related to the comparison between the predicted consumption and the real consumption during the last month:

- Consumption lower than the predicted by 8% or more
- Consumption lower than the predicted by 3 to 8%
- Consumption similar to the predicted (between 0 and 3% deviation)
- Consumption higher than the predicted by 3 to 8%
- Consumption higher than the predicted by 8% or more

By clicking on a building name from the Building List, a page will be open with information about the selected building.

In the building dashboard the user can access a recap of the registered energy efficiency measures, the total investment and the total electricity and heat consumption and savings.

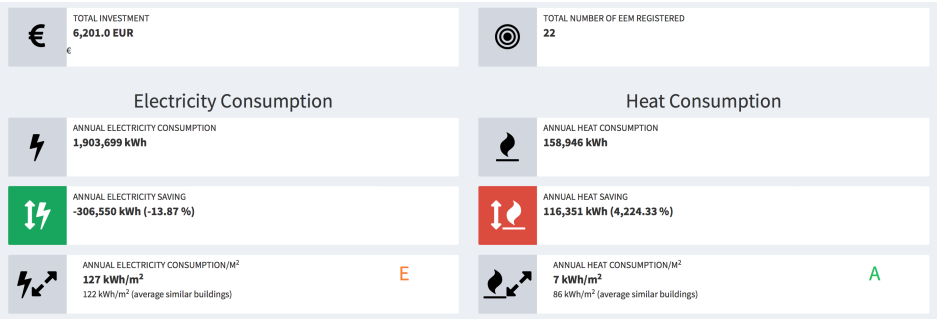

Figure 2.4: Building dashboard in the EDI-Net platform

In the same page, the user can access graphs representing the trend of the electricity and heat consumption for the last year. Two lines are plotted: the baseline (predicted) consumption and the actual consumption.

If the user uploaded hourly data, he'll be able to access more detailed graphs, that include hourly heat and electricity consumption. We can see an example of hourly data visualization

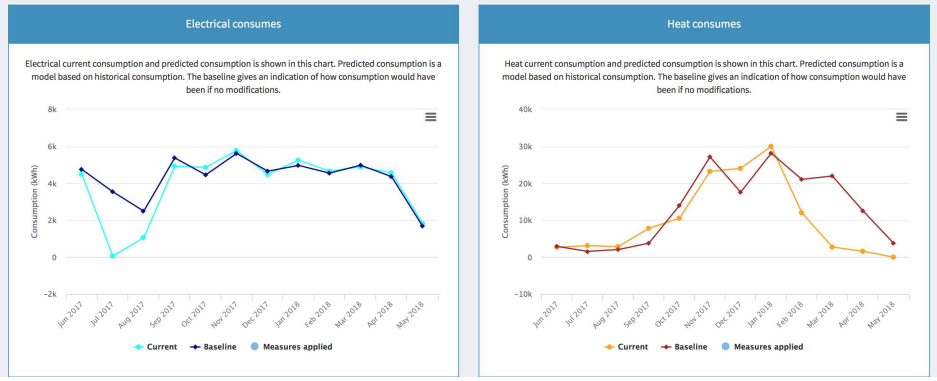

Figure 2.5: Actual and predicted consumption graphs in the EDI-Net platform

in Figure 6 below: the black line in the graph represents the actual consumption, while the colored bands help users to understand how close the actual consumption is to the hourly predicted consumption. If the black line coincides with the center of the yellow band, it means that the actual consumption is equal to the predicted one, if the black line is in the green (red) area, the consumption is lower (higher) than then predicted.

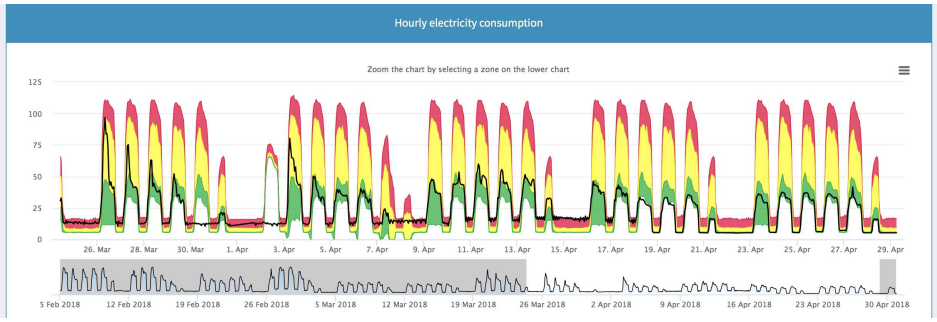

Figure 2.6: Hourly electricity consumption graph in the EDI-Net platform

Finally, users can access electricity and heat consumption comparison between the selected building and similar buildings.

If information about the total area of the building was added, the user can decide to visualize on the graphs either the absolute values or the surface values. The graph will show how the selected building is performing compared to other similar buildings and to the most efficient buildings of the same typology, as well as a letter representing the efficiency of the building (kWh/m$\hat{2}$y).

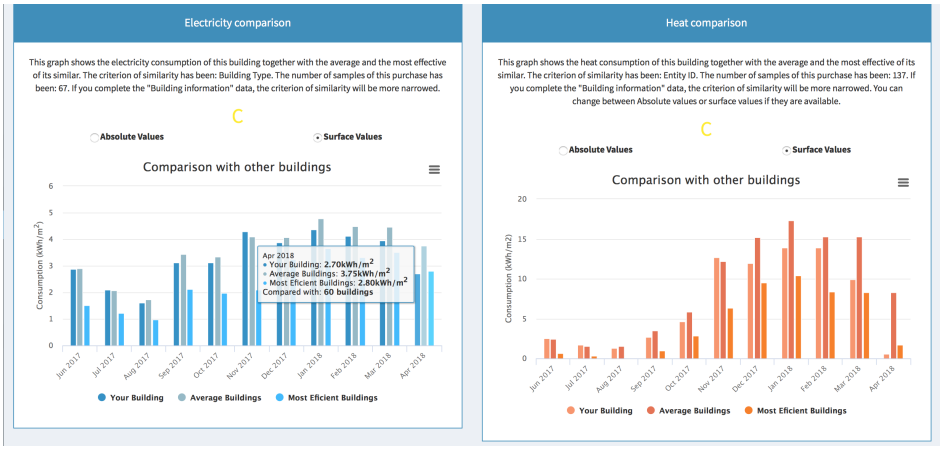

Figure 2.7: Building energy performance comparison

Chapter 3

Data Science concepts and state of the art

In this section, some statistics and data science concepts that will be used in the thesis are briefly introduced, together with a recap of state of the art techniques for energy savings evaluation and an introduction to the most common algorithms used for energy consumption data mining.

3.1 Statistic concepts

Residuals, R^2, R^2 adjusted and p-values are introduced here, as they are the main tools that will be used to evaluate the accuracy of a given model and the significance level of the different variables.

3.1.1 Residuals and RSS

In statistics the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest [9]. If we consider a dataset with n values, represented by the vector:

$$Y = [y_i, ..., y_n]^T \tag{3.1}$$

and each of these points is associated with a predicted value, calculated through a given statistical model:

$$\hat{Y} = [\hat{y}_i, ..., \hat{y}_n]^T \tag{3.2}$$

then we can define the ith residual (the difference between the ith observed response value and the ith response value that is predicted by our statistical model) as follows:

$$e_i = y_i - \hat{y}_i \tag{3.3}$$

We then define the *residual sum of squares* (RSS) as:

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 = (y_1 - \hat{y}_1)^2 + ... + (y_n - \hat{y}_n)^2 \tag{3.4}$$

At the same time, if we call $\bar{y}$ the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.5}$$

we can then define the *total sum of squares* (TSS) as:

$$TSS = \sum_i (y_i - \bar{y})^2 \tag{3.6}$$

3.1.2 Coefficient of determination (R^2)

A coefficient that provides a measure of how well a given model is able to replicated observed data, based on the proportion of total variation of outcomes explained by the model.

If we consider the same dataset of n values introduced previously:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.7}$$

TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression (or any other statistical model) is performed. In contrast, RSS measures the amount of variability that is left unexplained after performing the regression. Hence, $TSS - RSS$ measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using the statistical model. An R^2 that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response [9].

3.1.3 Adjusted R^2

The use of adjusted R^2 is an attempt to take account of the phenomenon that as new explanatory variables are added to the model, R^2 either maintains the same value or increases. To solve this issue, a penalizing factor is introduced, that helps identify the variables that are just adding noise to the model. Adjusted R^2 is defined as follows:

$$R^2 \; adjusted = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{3.8}$$

where p is the total number of explanatory variables in the model, and n is the sample size.

The intuition behind the adjusted R^2 is that once all of the correct variables have been included in the model, adding additional noise variables will lead to only a very small decrease in RSS. Since adding noise variables leads to an increase in p, such variables will lead to an increase in $\frac{RSS}{n-d-1}$ and consequently a decrease in the adjusted R^2. Therefore, in theory, the model with the largest adjusted R^2 will have only correct variables and no noise variables [9].

3.1.4 P-value

The p-value is a concept used in statistical hypothesis testing. Roughly speaking. it can be interpreted as the probability to observe a substantial association between a certain predictor and the response due to chance, in the absence of any real association between the predictor and the response. If the p-value, for a given predictor, is small (usually less than 0.05), then we can infer that there is an association between the predictor and the response. We then reject the null hypothesis, meaning that we declare a relationship to exist between the predictor variable X and the response Y. Maybe add the t value explanation as well if we need to put mathematical explanations [9].

3.1.5 Autoregressive model

In statistics, an autoregressive (AR) model is a representation of a type of random process. It is used to describe certain time-varying processes, while specifying that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term). Simple autoregressive models can be represented by the following equation:

$$X_t = \sum_{i=1}^{n} \phi_i X_{t-i} + \epsilon_t \tag{3.9}$$

where ϕ_i are the auto-regression coefficients, X_t is the series under investigation, n is the order of the autoregressive model and ϵ_t is the noise term, which is almost always assumed to be Gaussian white noise [**?**].

3.1.6 Clustering and K-means algorithm

Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set. When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other. Of course, to make this concrete, we must define what it means for two or more observations to be similar or different. Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied [9].

K-means clustering is a simple and elegant approach for partitioning a data set into K distinct, non-overlapping clusters. It is one of the oldest and best known clustering algorithms and it is still widely used for different purposes, often in combination with other clustering methods, such as Self-Organizing Maps. It attempts to group the observations of a data set into a fixed number of clusters with a minimized within-cluster variance. In the current case, the algorithm of Hartigan and Wong (1979) was chosen, which uses Euclidean distance to calculate the variance and defines the cluster center as the mean of each dimension of all observations in the cluster.
K-means is a relatively simple and fast algorithm, but it has a number of drawbacks, that have to be kept in mind when applying it. One of them is that it should be taken into account that the result of the algorithm depends on the randomly initiated cluster centres. Therefore, the algorithm is executed multiple number of times with different initializations to assure convergence.

3.1.7 Silhouette analysis

Silhouette analysis is a technique that can be used to evaluate the quality of a clustering. It is based on the calculation of the silhouette coefficient, through the following steps:

1. Calculate the cluster cohesion $a^{(i)}$ as the average distance between a sample $x^{(i)}$ and all other points in the same cluster.
2. Calculate the cluster separation $b^{(i)}$ from the next closest cluster as the average distance between the sample $x^{(i)}$ and all samples in the nearest cluster.
3. Calculate the silhouette $s^{(i)}$ as the difference between cluster cohesion and separation divided by the greater of the two, as shown here:

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{max\{b^{(i)}, a^{(i)}\}} \tag{3.10}$$

The silhouette coefficient is bounded in the range -1 to 1. Based on the preceding formula, we can see that the silhouette coefficient is 0 if the cluster separation and cohesion are equal ($b^{(i)} = a^{(i)}$). Furthermore, we get close to an ideal silhouette coefficient of 1 if $b^{(i)} >> a^{(i)}$, since $b^{(i)}$ quantifies how dissimilar a sample is to other clusters, and $a^{(i)}$ tells us how similar it is to the other samples in its own cluster, respectively [10].

3.2 State of the art of energy performance assessment

Assessing energy savings achieved after energy efficiency improvement measures might prove very challenging, especially when a combination of measures having cross effects are applied. Building owners of many large commercial buildings have ESCOs (Energy Service Companies) conduct energy audits to assess the most effective energy retrofits and plan management strategies. Although, detailed energy audits involve a complex process of data collection over long time durations, the development of an energy model, and several simulations for an accurate analysis. [11]

One of the approaches to solve this issue is the use of energy performance simulation tools. Such is the case of DEEP, a database of energy efficiency performance developed in California using the results of about ten million EnergyPlus simulations [12]. Other approaches use results of energy simulation tools and attempt to perform a correction with climatic and consumption measured data [13]

This thesis describes an innovative approach to evaluate energy savings without use of any energy simulation tool. The approach is based on the IPMVP (International Performance Measurement and Verification Protocol) Framework and, more specifically, on Option C of the protocol. Option C of the IPMVP involves use of utility meters to assess the energy performance of a total facility through a regression model or other data mining algorithm [14]

The algorithms most commonly used in data mining for the purpose of characterizing building energy consumption are prediction models. The two main classes are black box models and grey box models.
The class of predictive models that are difficult to interpret in terms of the drivers of energy use can be labelled as black box models. Their unique objective is prediction accuracy, and the mechanism responsible for their predictions contains little information about the system being modelled. Thus, these techniques, are only suitable for prediction models designed for forecasting purposes. The coefficients of the models lack of physical significance.
Models based on both insight into the system and experimental data are called grey box models. The parameter fit of these models yields meaningful physical information about the systems. Those whose primary purpose is to deliver meaningful parameter fits are called inverse models because they work backward from observations to reconstruct system parameters. They are useful to recover and provide semi-physical information from residential smart meter data.
Other data mining techniques useful to infer hidden information from consumption time series would be clustering and deep learning algorithms.

Linear steady state models

This technique determines the linear relationship between the consumption data and climate during a defined period. Nowadays, these low fine-grained models are already used in order to make a first estimation of weather dependence in consumption. [15] uses linear regression models to fit monthly heating fuel consumption, assuming a baseline of consumption not climate dependent α, and a slope β which represents the response of the building to outside temperatures.

The data requirements are:

- Daily / monthly consumption data.
- Daily outdoor temperature data or monthly HDD/CDD.

Piecewise linear models

Changepoint temperature models are based on this technique, a piecewise linear model between daily consumption and temperature response, with one or two breakpoints, corresponding to heating and cooling thresholds.
The response to outside temperature therefore has up to three segments. For two segment cases, one of the segments can be fixed to zero thermal response to capture temperatures below the cooling threshold or above the heating threshold without conditioning. Alternately, the temperature response of both segments can be fit by the model to capture both heating and cooling. The three-segment model consists of heating and cooling segments separated by a zero-slope segment for temperatures without conditioning. [16] describe a set of standardized regression tools for fitting interval-meter data.

The data requirements are:

- Daily consumption data.
- Daily outdoor temperature data.

Locally Weighted models

This method, described in [17], gives local estimates in time of the model coefficients by only considering observations within a limited time window. This makes it possible to see if they are constant over time, e.g. to look for variations during the heating season and how they change during the summer period. They are able to estimate the energy performance of buildings based on daily consumption measurements and nearby climate measurements.

The data requirements are:

- Daily consumption data.
- Daily weather data.

Autoregressive models

These techniques can be used for linear and stationary, not time-varying dynamical systems. However, in some cases a non-linear transformation of the input signals might be sufficient for the use of autoregressive models.
Depending on the application and the properties of the building/dwelling an appropriate sampling time range from e.g. five minutes to an hour should be considered.

Multiple implementations of this technique are described in [18].

The data requirements are:

- Hourly/Subhourly consumption data.
- Hourly/Subhourly weather data.

3.3 Thesis Methodology

The thesis methodology displayed in the flowchart in Figure 3.1. First, the data collection and cleaning phase are discussed, and the data model presented. Then, the different pilot buildings analyzed in this thesis are introduced, together with the five models developed to evaluate the energy performance. Finally, the different approaches for savings calculation are shown, and the algorithm results for the different pilots discussed.

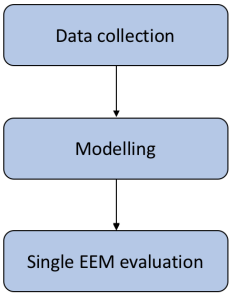

Figure 3.1: Thesis methodology

Chapter 4

Data collection

In this section, the data collection process is presented: first the data flux is described, providing an explanation of how the different quantities needed for the analysis are downloaded and merged in a single dataframe. Then, the process of evaluation eligibility is described and schematized, and the use of dummy variables is introduced. Finally, the data model is discussed with some examples.

4.1 Data flux

The data flux, represented in Figure 4.1 in the form of a flowchart, provides the reader with an understanding of the data source and explains the functions used to transfer the data from the database to the final dataframe used for the analysis described in this thesis.

The first step is the creation of a function that allows to download from MongoDB information about all the buildings that are eligible for the analysis, that is to say buildings that applied energy efficiency measures since January 2016 and that have hourly consumption data. Through this function we can access:

- the details about the applied EEM (category and application date),
- the modelling unit ID (a unique string needed to access the consumption data of the selected building),
- latitude and longitude of the building location, needed to download the weather data.

Two more functions allow us to download the desired consumption values from MongoDB and to get historical weather stations observations from the Darksky API (https://darksky.net/dev). The consumption data can be either on a 30 minutes or on a hourly base, the weather data is sampled every hour by the closest weather station to the GPS coordinates of the selected building and includes information about temperature, solar irradiance, sun elevation, relative humidity, wind speed and wind direction.

After having downloaded the consumption data, the EEM details and the weather observations, everything is merged into one single dataset. Since the models we developed work

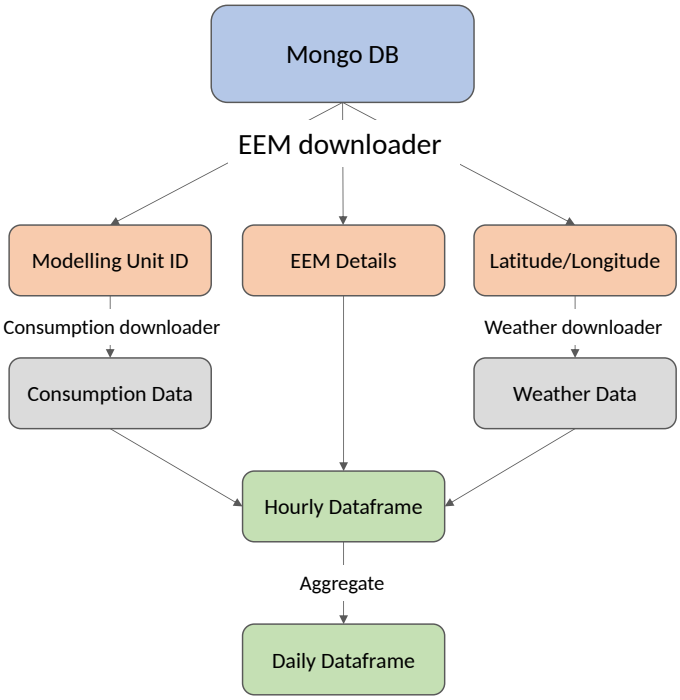

Figure 4.1: Data flux scheme

with daily datasets, a daily aggregation is performed on the hourly dataset to create the final dataframe, ready to be analyzed.

4.2 Decision tree and dummy variables

Once the information is merged in a single dataset and the daily aggregation is performed, the applied EEMs are analyzed, in order to understand whether the impact evaluation is possible or not. This process is schematized in the decision tree in Figure 4.2.

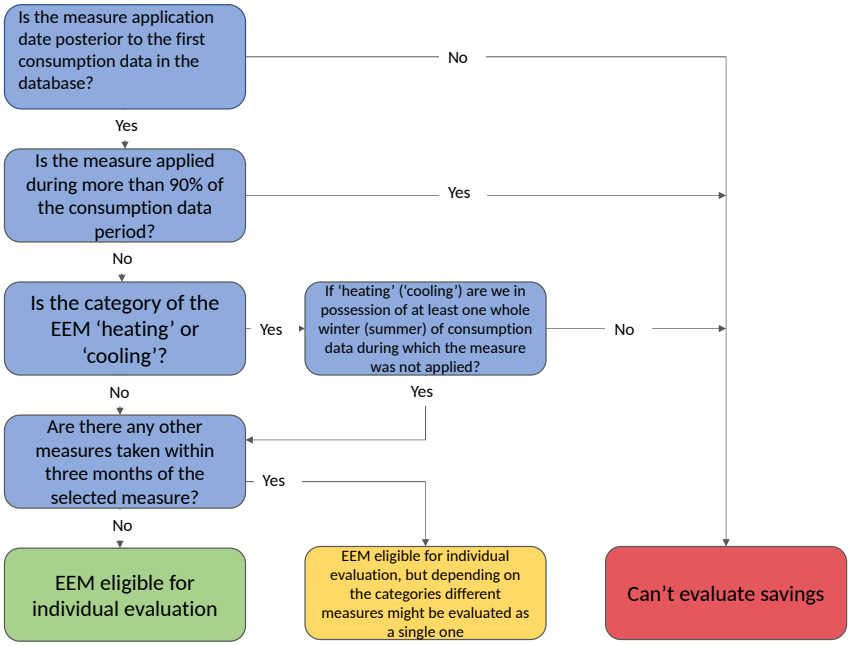

Figure 4.2: Evaluation feasibility decision tree

This decision tree is based on the fact that, in order to evaluate an EEM, training data for the model is needed, that is to say the model must have access to a certain amount of data, prior to the EEM application, to be able to detect possible changes in the consumption pattern. This means that, in order to be eligible for evaluation, a measure must not be applied for more than 90% of the consumption timeseries we are analyzing. Furthermore, if the measure belongs to the category 'heating' (or 'cooling'), an additional requirement is added: at least one winter (or summer) of data without the measure applied is required. Finally, some additional exceptions are added in case there are multiple different EEMs applied and their application dates are close one to the other, depending on how close the measures are, it might be impossible to evaluate them separately.

For every measure that is eligible for evaluation, a dummy variable m is added to the dataset, having value 0 before the measure application date and value 1 after the measure application date. m is essential to differentiate the section of the timeseries where the EEM is not applied and the one where it is, it can be described as:

$$m = \begin{cases} 0 & \text{if the measure is not applied} \\ 1 & \text{if the measure is applied} \end{cases} \tag{4.1}$$

To be able to better evaluate the impact of 'heating' and 'cooling' measures, which are supposed to generate savings only during specific months, the dummy variables for these categories were set to 1 only during the heating and cooling months following the measure application date. Heating months were considered to be December, January and February, cooling months were considered to be June, July and August.

4.3 Data cleaning

The first step that should be followed when analysing massive datasets is to clean the time series. Data cleaning in the energy domain is a process of detecting, diagnosing, and editing faulty data in consumption time series. In the methodology developed, this is a critical step because the analysis is made based on the raw electricity consumption time series. As shown in [19], there are multiple approaches to detect outliers in consumption time series.

In this study, two outliers detection approaches are considered. The first is a non-recursive elimination of extreme scores based on a Z-score of one-week sliding window population in order to detect outliers when their value is above eight (it means eight standard deviations from the mean). The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being measured. The second method considers a measure as an outlier when its value is above the theoretical maximum consumption based on the customer contracted power.

Additionally, data gaps are detected using a data-padding algorithm, which already considers the original time series frequency. Not Available (NA) values are the replacement for each of the faulty timestamps in raw time series.

4.4 Data model

The structure of the dataframes used in the algorithm is briefly presented in this section, an example is shown in Figure 4.3.

Timestamp	E_d	T_e	Daylight hours	Structure	Weekday	m1	AR1	...	AR7
2016-01-10	140.41	4.48	6	4	7	0	226.9	...	84.07
2016-01-11	219.70	3.62	6	1	1	0	140.4	...	84.98
2016-01-12	223.16	4.61	6	1	2	0	219.7	...	161.8

Figure 4.3: Example of dataframe structure

The information contained in the dataframe:

Timestamp: date in format YYYY-MM-DD

E_d: electricity consumption of the day in kWh

T_e: daily average of the external temperature

Daylight hours: number of hours of light of the day

Structure: a variable indicating the consumption pattern of the building, obtained through the clustering algorithm presented in section 5.4

Weekday: day of the week (1 = Monday, 2 = Tuesday, etc.)

m1: dummy variable indicating if a given measure is applied or not

AR1 ... AR7: autoregressive variables indicating the consumption of 1 to 7 days before

Chapter 5

Modelling

The objective of this section is to present the different models and approaches that were evaluated with the goal of assessing the impact of a single energy efficiency measure. Three different pilot buildings were analyzed, their characteristics are briefly presented in the following paragraphs, together with the models that were tested.

The main approach that was chosen was to develop a statistical model able to fit correctly the consumption and weather data and to assess the impact of the measure by evaluating how the model would differently fit the data before and after the EEM implementation. Among the different possible statistical models, it was decided to work with generalized additive models (GAM), a specific method for supervised learning, originally developed by statisticians Trevor Hastie and Robert Tibshirani. All the models described in this section were realized using R, a programming language widely used for data analysis and statistical computing.

5.1 Generalized Additive Models

Although attractively simple, the traditional linear model often fails in many situations, since in real life effects are often not linear. GAMs are flexible statistical methods that can be used to identify and characterize nonlinear regression effects. In the regression setting, a generalized additive model has the form:

$$g(E(Y)) = \beta_0 + f_1(X_1) + f_2(X_2) + ... + f_p(X_p) \tag{5.1}$$

where $X_1, X_2, ..., X_p$ represent the predictors and Y is the outcome; the f_j's may be functions with a specified parametric form (polynomial or un-penalized regression spline, for example), or unspecified 'smooth' functions, to be estimated by non-parametric means. This means that the model allows for rather flexible specification of the dependence of the response on the covariates, but this flexibility comes at the cost of two theoretical problems: it is necessary to represent the smooth functions and to choose how smooth they should be.

Our approach is to fit each function using a scatterplot smoother and provide an algorithm that simultaneously estimates all p functions. The scatterplot smoother we used in our analysis is a cubic smoothing spline.

5.2 Pilot 1: Seu Central del Departament d'Interior (ES)

This pilot is made of a complex of five buildings, managed by 'Generalitat de Catalunya' and located in the city of Barcelona. A recap of parameters useful for the performed analysis follows:

- Total building area: 19 131 m^2
- Building year: 1947
- Schedule: Monday-Friday 7:00-22:00
- Annual electricity consumption: 2 756 638 kWh
- Annual electricity consumption/m^2: 110 kWh/m^2
- Heating main source: electricity

As we can see the building has a very large area and a substantial electricity consumption. For the testing and validation of the different models, hourly electricity consumption data starting 2016-01-01 was used, as for the EEMs, the last three measures registered in the EDINET platform were considered, as presented in Table 5.1.

Date	Source	Description	Investment (€)
01/06/16	Lighting	General	6 200
01/05/16	Cooling	Complete replacement of the cooling equipment	–
01/05/16	Heating	Replacement of heat production equipment with a more efficient one	300 000

Table 5.1: EEMs for Seu Central del Departament d'Interior

As a preliminary step for the analysis, the electricity consumption timeseries of the building was analysed, together with the temperature data, in order to evaluate its weather dependency. The two graphs are shown in Figure 5.2 and show a clear weather dependency of the

Figure 5.1: Seu central del Departament d'Interior: view from outside

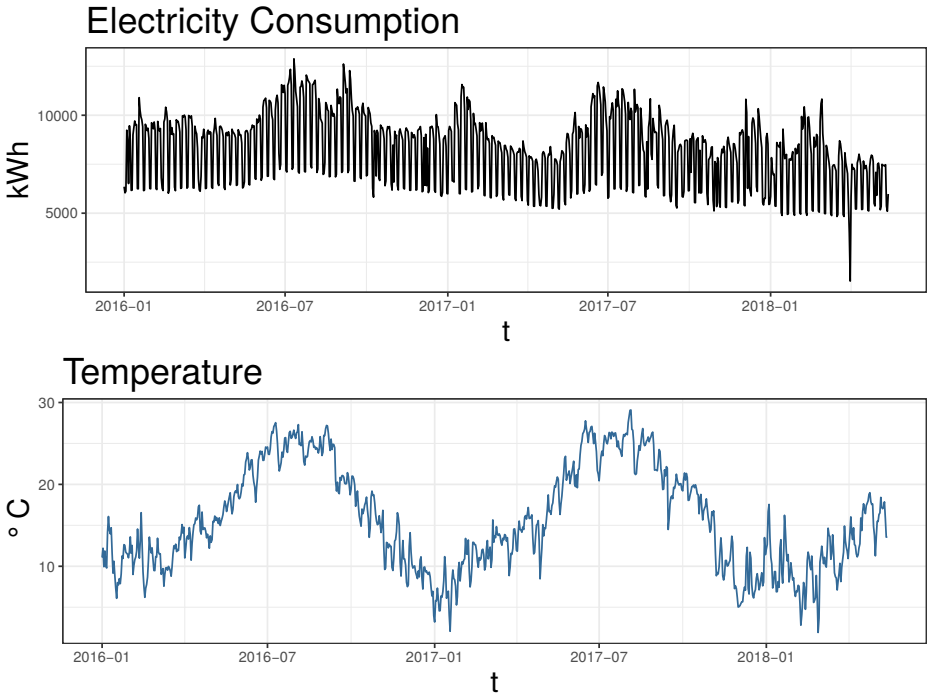

Figure 5.2: Electricity consumption and external temperature graphs for Seu central del Departament d'Interior

building, that reaches its consumption peaks, both in summer and in winter, in conjunction with the highest and lowest external temperatures.

By analysing the hourly electricity consumption pattern over any week, it is also possible to notice the typical 'office building' behaviour, with peaks during the day and low consumption at night and during the weekend.

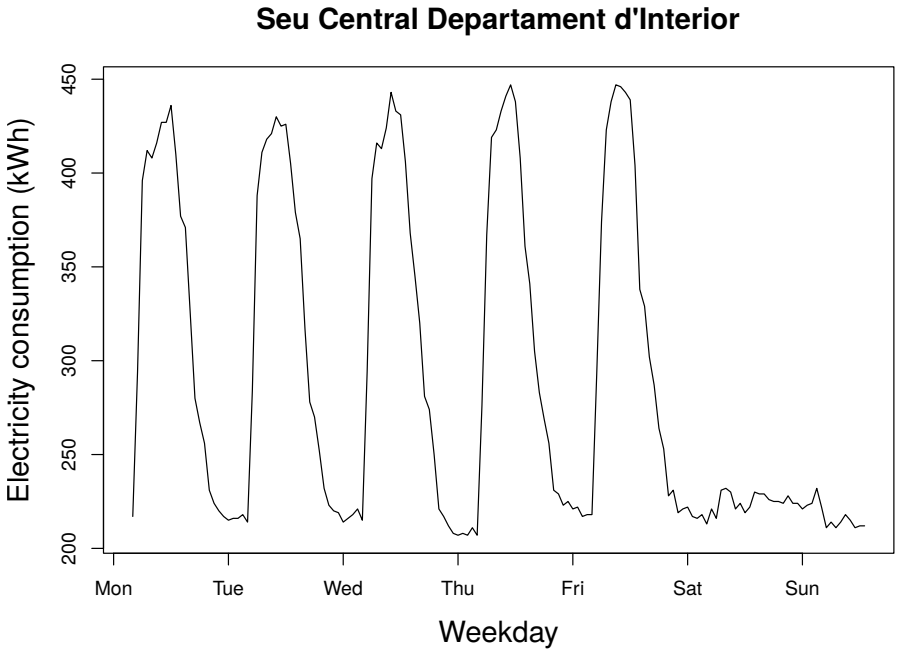

Figure 5.3: Electricity consumption weekly pattern for Seu Central del Departament d'Interior

Two different models were tested for these pilot, in both cases the heating and the lighting measures were the only ones considered. The cooling measure was discarded because it was applied in May 2016, but the consumption data we have from this building starts in January 2016. This means there is no data available to analyze the building consumption pattern during summer, before the application of the measure. For this reason, the cooling measure was discarded and the model attempts to evaluate exclusively the lighting and the heating measure.

Model 1.1 - GAM with smooth functions and linear measure predictors

The first model that was developed was a GAM having two smooth functions with temperature and day of the week and a linear predictor per every energy efficiency measure (one for the lighting measure and one for the heating measure). The formula used is the following:

$$E_d = B_l + f_1(T_e) + f_2(d_w) + \alpha_1 m_1 + \alpha_2 m_2 \tag{5.2}$$

where

E_d is the daily electricity consumption

B_l is the base load of the building, that is independent from external parameters.

f_1 and f_2 are two 'thin plate' smooth functions (this needs to be described somewhere)

T_e is the external temperature (daily average)

d_w is a parameter that represents the day of the week (1 for Monday, 2 for Tuesday etc.)

m_1 and m_2 are coefficients that are equal to zero when the corresponding measure is not applied and equal to one when it is applied.

Thanks to the coefficients m_1 and m_2 we can make sure that the period prior to the measure application is considered as a training period for the model, and the usual behaviour of the building (prior to the application of the measures) can be then represented by the factors $B_l + f_1(T_e) + f_2(d_w)$, while the impact of the two measures will be contained in the terms $\alpha_1 m_1 + \alpha_2 m_2$.

The model described is very convenient, as it provides a way to quickly evaluate the effect of the measures, since the impact is represented by the different coefficients α and there is no need for further calculations.

Model 1.2 - GAM with autoregressive variables and temperature splines

This model is an autoregressive model that attempts to evaluate the electricity consumption pattern using as variables the shifted daily consumption values and smooth functions representing temperature dependency before and after the heating measure; the lighting measure is evaluated through a linear coefficient. The model can be described as follows:

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i \omega_i) + f_1(T_e)_{[m_1=0]} + f_2(T_e)_{[m_1=1]} + \alpha m_2 \tag{5.3}$$

where:

B_l is the base load of the building,

$\sum_{i=1}^{7}(\phi_i \omega_i)$ are the seven daily consumption auto-regressive variables,

$f_1(T_e)_{[m_1=0]}$ and $f_2(T_e)_{[m_1=1]}$ are the two smooth functions representing the relation between consumption and temperature before and after the EEM application,

αm_2 is the linear term representing the impact of the lighting measure.

Although its interpretation might be less intuitive and involve some further calculations, this model surely describes better the dynamic of the building consumption pattern. Moreover, thanks to the two temperature splines we are able to detect with increased accuracy the impact of the heating measure. It's important to point out that the temperature smooth function after the measure application $f_2(T_e)_{[m_1=1]}$ was only fitted using winter months, in order to avoid any additional noise.

5.3 Pilot 2: Highfields Library (UK)

This building is managed by the Leicester City Council and is located in the city of Leicester (UK). A recap of parameters useful for the performed analysis follows, an external view of the building is shown in Figure 5.4:

- Total building area: 506 m^2
- Schedule: Monday-Friday 10:00-19:00, Saturday 10:00-16:00
- Annual electricity consumption: 38 440 kWh
- Annual electricity consumption/m^2: 69 kWh/m^2

Figure 5.4: Highfields library: view from outside

For the testing and validation of the model, electricity consumption data sampled every 30 minutes and starting 2016-01-01 was used. Since 2016, only one EEM was applied, its details are described in Table 5.3.

Date	Source	Description	Investment (€)
27/04/18	Lighting	General	12 083

Table 5.2: EEMs for Highfields Library

In Figure 5.5 we can see the electricity consumption and temperature graphs for Highfields Library. The pattern suggests that the electricity consumption does not have weather dependency, as the values are very similar throughout the year and do not seem to have any relation with the external temperature. It is also possible to identify a period of irregular functioning, during the month of April 2017, possibly indicating a week during which the building was closed, due to maintenance or some other unknown reasons.

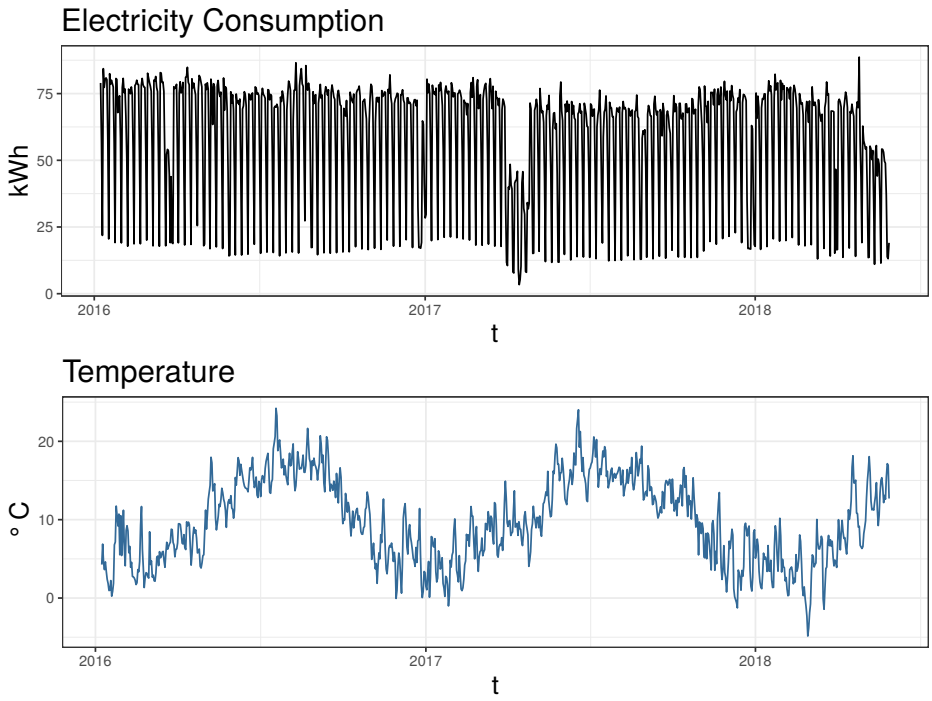

Figure 5.5: Electricity consumption and external temperature graphs for Highfields Library

By analyzing the weekly profile in Figure 5.6 we can clearly see the building schedule that was mentioned at the beginning of the paragraph. As the building's electricity consumption does not have any weather dependency, we can conclude that we are mainly looking at lighting and electrical appliances consumption, this is also confirmed by the fact that the values are quite low (hourly peaks do not reach 5 kWh). It is also interesting to see that during the weekdays there is a first consumption peak that is reached between 05:00 and 06:00 and that probably corresponds with the cleaning schedule of the building.

Model 2.1 - GAM with autoregressive variables and linear measure predictors

This model uses autoregressive variables to describe the general consumption pattern of the building and a linear coefficient to detect the impact of the EEM:

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i \omega_i) + \alpha m_1 \tag{5.4}$$

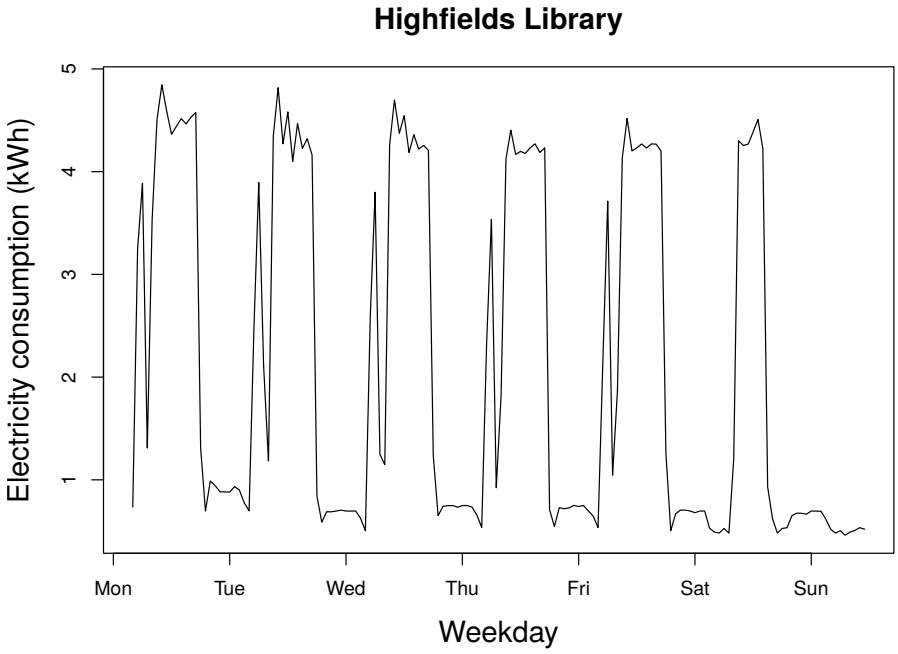

Figure 5.6: Electricity consumption weekly pattern for Highfields Library

where:

B_l is the base load of the building,

$\sum_{i=1}^{7}(\phi_i \omega_i)$ are the seven daily consumption auto-regressive variables,

$\alpha_8 m_1$ is the linear term representing the impact of the lighting measure.

The model is similar to the autoregressive GAM introduced previously for the Spanish pilot. The main difference is that the measure applied here belongs to the category "lighting", therefore its effect was supposed to be independent from the temperature variable and the two temperature smooth functions were taken out of the GAM equation. For a more accurate prediction, smooth function representing the dependency of the building consumption on the hours of light of a given day might be added. This was not done here because the application date of the EEM is 27/04/2018 and it was supposed that one month of data would not be enough to properly train the smooth functions.

5.4 Pilot 3: Belgrave Neighbourhood Centre (UK)

This building is managed by the Leicester City Council and is located in the city of Leicester (UK). A recap of parameters useful for the performed analysis follows, an external view of

the building is shown in Figure 5.7:

- Total building area: 1699 m^2
- Schedule: Monday-Saturday 09:00-22:00 Sunday 08:00-16:00
- Annual electricity consumption: 78 744 kWh
- Annual electricity consumption/m^2: 41 kWh/m^2
- Heating main source: natural gas

Figure 5.7: Belgrave Neighbourhood Centre: view from outside

For the testing and validation of the model, electricity consumption data sampled every 30 minutes and starting 2016-01-01 was used. Since 2016, only one EEM was applied, its details are described in Table 5.3.

Date	Source	Description	Investment (€)
30/09/17	Lighting	General	17 529

Table 5.3: EEMs for Belgrave Neighborhood Centre

In Figure 5.8 it is possible to see that the building reaches higher consumption levels during winters than during summers, although during winter 2017/18 the consumption was not so high as during the two previous. This might be due to the Lighting EEM that was taken in Autumn 2017.

The analysis of the weekly consumption pattern suggests that, unlike the previously introduced pilots, in this building daily consumption patterns might be very different from day to day. This is why, in order to obtain an accurate model, it was necessary to apply a clustering algorithm, as presented further in this paragraph.

Model 3.1 - GAM with autoregressive variables and daylight hours splines

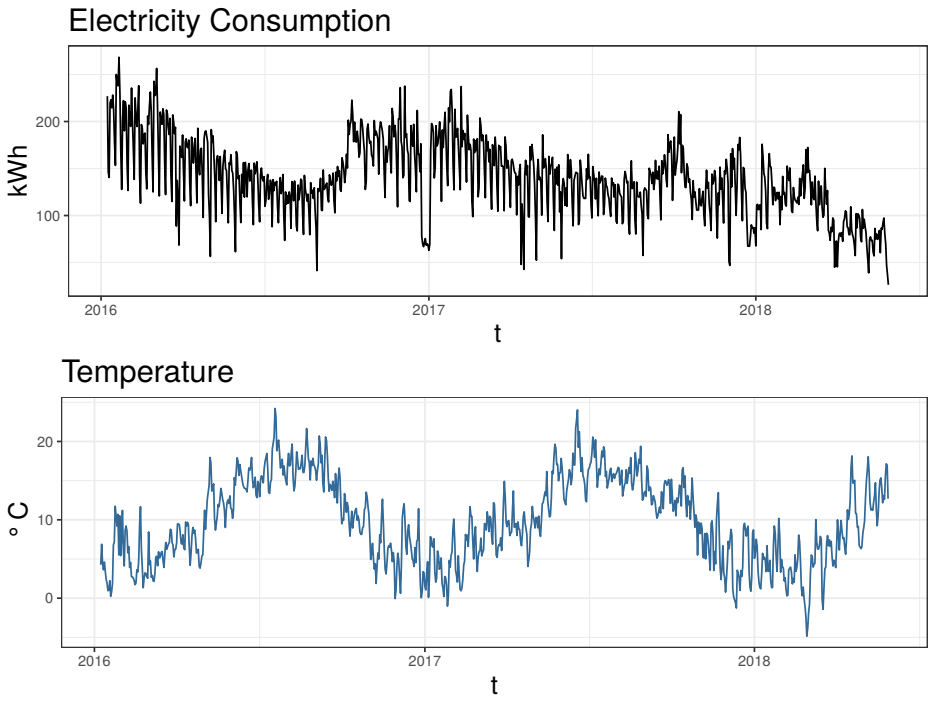

Figure 5.8: Electricity consumption and external temperature graphs for Belgrave Neighbourhood Centre

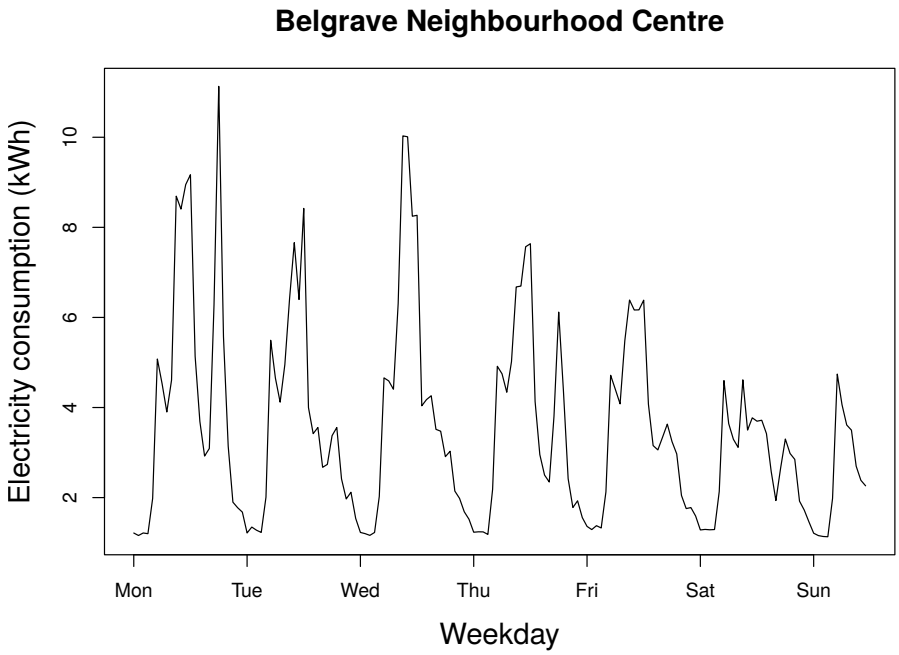

Figure 5.9: Electricity consumption weekly pattern for Belgrave Neighbourhood Centre

Similarly to model 2.2, which was previously introduced, this model uses autoregressive variables to describe the general consumption pattern of the building and two smooth functions of the daylight hours variable, to evaluate the impact of the lighting EEM:

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i \omega_i) + f_1(d_h)_{[m_1=0]} + f_2(d_h)_{[m_1=1]} \tag{5.5}$$

where:

B_l is the base load of the building,

$\sum_{i=1}^{7}(\phi_i \omega_i)$ are the seven daily consumption auto-regressive variables (and associated linear coefficients),

$f_1(d_h)_{[m_1=0]}$ and $f_2(d_h)_{[m_1=1]}$ are the two smooth functions representing the relation between consumption and daylight hours before and after the EEM application.

This model was applied hoping to have similar results to the ones of model 2.2. Although, the R^2 adjusted and the significance levels of the splines proved to be quite low, as presented in the Results section. For this reason, an additional analysis, involving a K-means clustering algorithm, was performed on this pilot, in order to obtain a better fitting model and a more accurate evaluation of the EEM impact.

Model 3.2 - K-Means clustering, consumption pattern variables and linear measure predictor

The low R^2 adjusted shown by the previous model is a sign that the autoregressive variables alone are unable to properly represent the consumption pattern of the building. One of the reasons for this behaviour might be that, according to the specific day of the week or month of the year, the building follows a different consumption pattern, to which we will, from now on, refer as "structure".

Most research on time series analysis and forecasting is normally based on the assumption of no structural change on the time series, which implies that the variable model is stable and consistent during the whole data period. However, in building energy, time series data usually have different structures. For example, it is common to have a very different energy profile in winter days and in the summer days.

To detect the different structures in the data set of the selected pilot building, a clustering methodology was applied. A partitioning algorithm, based on K-means approach [20] and euclidean distance was used to identify the different k structures within the data set. The k-means model was used to classify the different days into groups of similar daily behaviour, using as input the building's daily consumption values, sampled every 30 minutes.

To decide the number of structures (k), a local maximum of the silhouette coefficient value is sought, with k being in the range [2,6]. This because the two structures case will most probably always have the highest silhouette coefficient, but at the same time considering only two structures would oversimplify the problem. A minimum threshold of 0.2 is defined, if not met, a single structure will be considered.

Once the optimal number of structures was calculated, through the silhouette analysis method, a GAM model with autoregressive variables, temperature splines and information about the daily structure was realized, with the following mathematical representation:

$$E_d = B_l + \sum_{i=1}^{7}\sum_{j=1}^{k}(\phi_{ij}\omega_{ij}) + \sum_{j=1}^{k} f_j(T_e) + \alpha m_1 \tag{5.6}$$

where

k is the optimal number of structures detected by the silhouette analysis,

$\sum_{i=1}^{7}\sum_{j=1}^{k}(\phi_{ij}\omega_{ij})$ are the $7 \times k$ autoregressive consumption variables (and associated linear coefficients) taking into account, per every variable, the structure of the day,

$\sum_{j=1}^{k} f_j(T_e)$ are the k temperature smooth functions taking into account the daily structure information,

α is the linear coefficient that represents the savings generated by the measure $m1$.

Chapter 6

Single EEM impact evaluation

In this paragraph, a brief summary of how the EEM savings are calculated, depending on the different models applied, is presented. The models introduced in the previous chapter can be divided into two categories, according to the two main approaches that were used to evaluate the EEM impact:

1. Linear coefficient evaluation (models 1.1,2.1,3.2)
2. Smooth function difference evaluation (models 1.2,3.1)

6.1 Linear coefficient evaluation

Although the three models belonging to this category have different ways of representing the standard consumption pattern (smooth functions, autoregressive variables, clustering structures), they all attempt to quantify the savings generated by the considered EEM through a coefficient α.

The mathematical equations representing these three models are reported here to allow an easier understanding of this section:

$$E_d = B_l + f_1(T_e) + f_2(d_w) + \alpha_1 m_1 + \alpha_2 m_2 \tag{6.1}$$

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i \omega_i) + \alpha m_1 \tag{6.2}$$

$$E_d = B_l + \sum_{i=1}^{7}\sum_{j=1}^{k}(\phi_{ij} \omega_{ij}) + \sum_{j=1}^{k} f_j(T_e) + \alpha m_1 \tag{6.3}$$

α is the linear coefficient that quantifies the effect of the dummy variables m_i, previously introduced in section 4.2. This means that, although defined as a linear coefficient, α is in reality a constant to be subtracted (or added) to the building standard consumption, during the period when the EEM is applied.

Since the models presented consider as predicted variable the daily electricity consumption E_d, α can be seen as the kWh saved per day, due to the EEM application. In order to better understand the impact of the measure, α needs to be compared to the average daily consumption of the building, we then define s as the average daily savings percentage, due to the EEM application:

$$s = \frac{\alpha}{\bar{E}_d} \cdot 100\ \% \tag{6.4}$$

where $\bar{E}_d$ represents the average daily electricity consumption.

Once we know s, it is easy to calculate total yearly savings S_y, by simply multiplying it by the average yearly consumption, denoted here with the symbol $\bar{E}_y$:

$$S_y = s\bar{E}_y \tag{6.5}$$

Although very attractive because of its simplicity and intuitiveness, the flaws of this model are related to its accuracy. In fact, if the consumption decreases after a given EEM, this model is not able to tell which part of this decrease is actually due to the EEM application and which part is caused by other independent variables. A second, more precise, approach is presented in the next paragraph.

6.2 Smooth function difference evaluation

To improve the precision of the impact evaluation, two different models were realized, using smooth functions of specific variables as savings predictors. The mathematical equations representing these two models are reported here to allow an easier understanding of this section:

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i\omega_i) + f_1(T_e)_{[m_1=0]} + f_2(T_e)_{[m_1=1]} \tag{6.6}$$

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i\omega_i) + f_1(d_h)_{[m_1=0]} + f_2(d_h)_{[m_1=1]} \tag{6.7}$$

The savings generated by the applied measures are calculated here as a difference between the two smooth functions representing the consumption's dependence on the selected variables before and after the application of the measure.

To calculate the savings, the best fitting polynomials to the two splines are estimated and the sum of the polynomial values is calculated over the period of time during which the measure was applied. The savings are then considered to be equal to the difference between the two sums (in formula 6.8 the external temperature is considered as a variable, but the formula doesn't change for the daylight hours case):

$$S = \sum_{i=k}^{n} f_{m0}(T_e(t_i)) - \sum_{i=k}^{n} f_{m1}(T_e(t_i)) \tag{6.8}$$

where:

we are considering a consumption timeseries made of n daily values and supposing that the measure application date corresponds to the kth value,

f_{m0} is the best fitting polynomial to the pre-application smooth function,

f_{m1} is the best fitting polynomial to the post-application smooth function,

$T_e(t_i)$ is the external temperature timeseries for the analysed building.

This approach provides an increased accuracy in detecting the savings, since it does not analyse the general consumption trend, but the consumption is instead analyzed in relation to the variables connected to the specific EEM (heating/cooling-external temperature, lighting-daylight hours etc.).

By comparing the savings with the average yearly consumption and by normalizing the value for the number of days the measure has been active, we can obtain the impact of the measure in terms of percentage:

$$s = \frac{S}{\bar{E}_y} \cdot \frac{365}{n-k} \tag{6.9}$$

Chapter 7

Results

In this chapter, the results of the analysis are discussed. The different models are compared in terms of accuracy and the savings are calculated according to the methodologies described in Chapter 6. Different graphs and tables are used to better explain the algorithm results.

7.1 Model 1.1

The mathematical equations representing this model is reported here to allow easier understanding of this section:

$$E_d = B_l + f_1(T_e) + f_2(d_w) + \alpha_1 m_1 + \alpha_2 m_2 \tag{7.1}$$

The summary of the model results follows, in terms of R^2 adj., deviance explained and $p - values$:

R^2 adj. $= 0.786$

Deviance explained $= 78.8\%$

Variable	p-value	Variable	p-value
$p(f_1)$	$< 2 \cdot 10^{-16}$	$p(f_2)$	$< 2 \cdot 10^{-16}$
$p(m1)$	0.00708	$p(m2)$	$< 2 \cdot 10^{-16}$

Table 7.1: P-values Table for Model 1.1

In Figure 7.1 the two thin plate smoothing functions f_1 and f_2 are displayed. The day week spline is reflecting the typical office building consumption pattern, with sustained consumption during working days and a steep decrease during the weekend. From the temperature spline we can appreciate what we already noticed by comparing the temperature and the

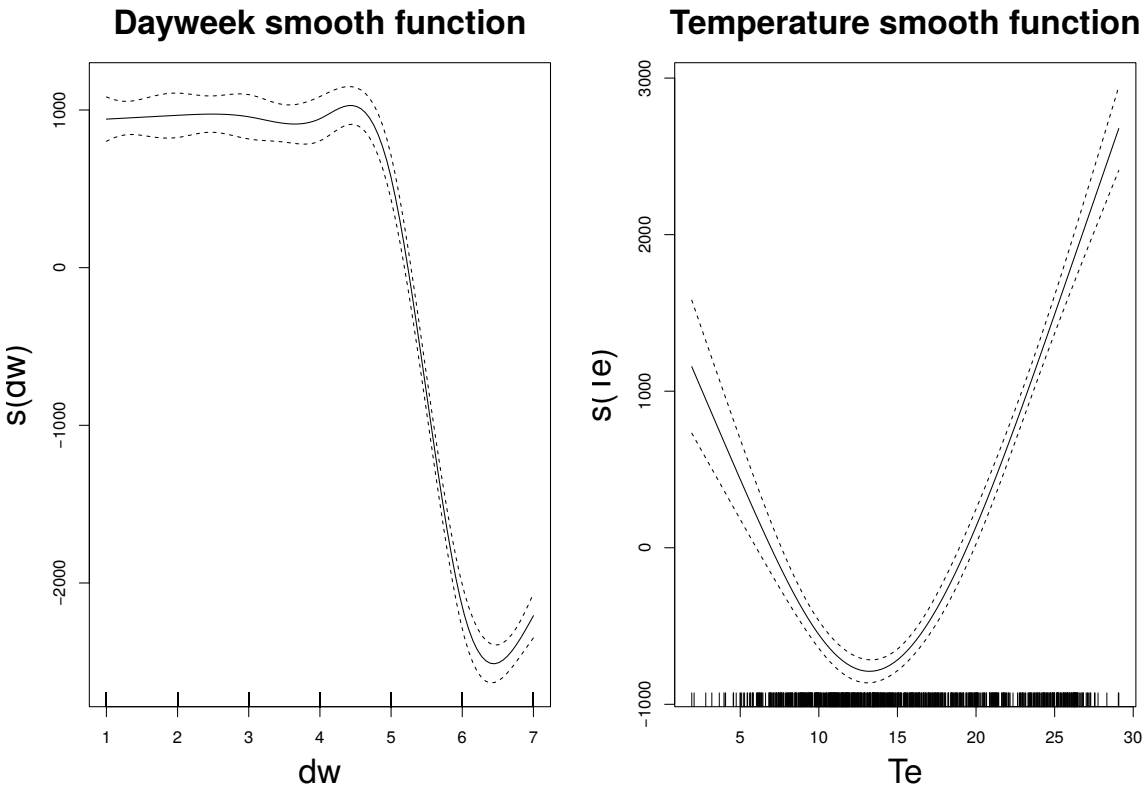

Figure 7.1: Dayweek and temperature splines for model 1.1

consumption timeseries: the building has weather dependency for both winter and summer, showing increased consumption for temperatures lower than 10 °C and higher than 15 °C.

The model fit assigned the following values to the coefficients α:

$\alpha_1 = 279.09$

$\alpha_2 = -957.71$

meaning that the lighting measure $m2$ generated savings of approximately 960 kWh per day, while the heating measure, $m1$, apparently caused an increase in the building consumption of about 280 kWh per day. We use now equation 6.4 to evaluate the impact of the two measures:

$$s_1 = \frac{\alpha}{\bar{E}_d} = \frac{279.09}{8232.9} = 0.034 = 3.4\ \% \tag{7.2}$$

$$s_2 = \frac{\alpha}{\bar{E}_d} = \frac{-957.71}{8232.9} = -0.12 = -12\ \% \tag{7.3}$$

Finally, we use equation 6.5 to evaluate the yearly kWh saved thanks to $m2$:

$$S_{m2-y} = s_2\ \bar{E}_y = 0.12 \cdot 2756638 = 330796.6\ kWh/year \tag{7.4}$$

According to this model, the heating measure caused an increase in the consumption, while the lighting measure was associated with substantial energy savings.

As we already mentioned previously, the linear coefficient estimation, although attractive for its simplicity, may in some cases provide incorrect results. In the case of this specific pilot, there are two important factors that might be influencing our model:

- the two measures' application dates are just one month apart one from the other,
- being the heating measure applied during summer, the savings that should come from this EEM are not immediate, but will only be appreciated several months after.

In order to get around these issues and improve the accuracy of the performance assessment, a model that uses GAM smooth functions to evaluate savings was built, its results are discussed in the next paragraph.

7.2 Model 1.2

The mathematical equation representing model 1.2 is reported here to allow easier understanding of this section:

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i \omega_i) + f_1(T_e)_{[m_1=0]} + f_2(T_e)_{[m_1=1]} \tag{7.5}$$

The summary of the model results follows, in terms of R^2 adj., deviance explained and $p-values$:

R^2 adj. = 0,796

Deviance explained = 79,9%

Variable	p-value	Variable	p-value
ω_1	$< 2 \cdot 10^{-16}$	ω_2	$2,64 \cdot 10^{-9}$
ω_3	0,36012	ω_4	0,70095
ω_5	$1,2 \cdot 10^{-6}$	ω_6	$3,89 \cdot 10^{-7}$
ω_7	$< 2 \cdot 10^{-16}$	$m2$	0.00217
$f_{m1=0}(T_e)$	$2,92 \cdot 10^{-14}$	$f_{m1=1}(T_e)$	$3,07 \cdot 10^{-6}$

Table 7.2: P-values Table for Model 1.2

From Table 7.2 we can see that all the variables, except ω_3 and ω_4, have very high significance levels.

As previously mentioned, this model attempts to evaluate the savings of $m2$ through a linear coefficient, while the savings of $m1$ are evaluated thanks to the use of two temperature smooth functions.

$\alpha_2 = -324,235$ kWh tells us that the lighting EEM caused savings of about 325 kWh per day, substantially lower than the ones obtained in Model 1.1. To calculate the savings generated from $m1$, the two temperature splines, here shown in Figure 7.2 have to be evaluated.

The graph shows lower consumption, for fixed temperature values, after the measure application. To calculate the exact total savings, the two best fitting polynomials to the curves were calculated. Second order polynomials provided an R^2 of 0.99, proving to fit almost perfectly the temperature splines. The two polynomials:

$P_0 = 1753.7995 - 179.876\ T_e + 5.2064\ T_e^2$

$P_1 = 1951.4785 - 255.974\ T_e + 8.331\ T_e^2$

To allow easier calculations, since we are interested in the absolute value of the difference between the two splines, the minimum value reached by P_1 (-417.1986 kWh) is summed to the values of both polynomials, in order to get rid of any negative value.Then, applying the R function *predict*, the sum of the polynomial values during the months with $m1 = 1$ is calculated, and the savings are evaluated as described in Section 6.2.

$$S_{m1} = \sum_{i=k}^{n} f_{m0}(T_e(t_i)) - \sum_{i=k}^{n} f_{m1}(T_e(t_i)) = 72557.77 - 109855.6 = -37297.86\ kWh \quad (7.6)$$

These are the total savings that the heating EEM generated during the two winters that followed its application (as it was explained in section 4.2 the dummy variables for heating

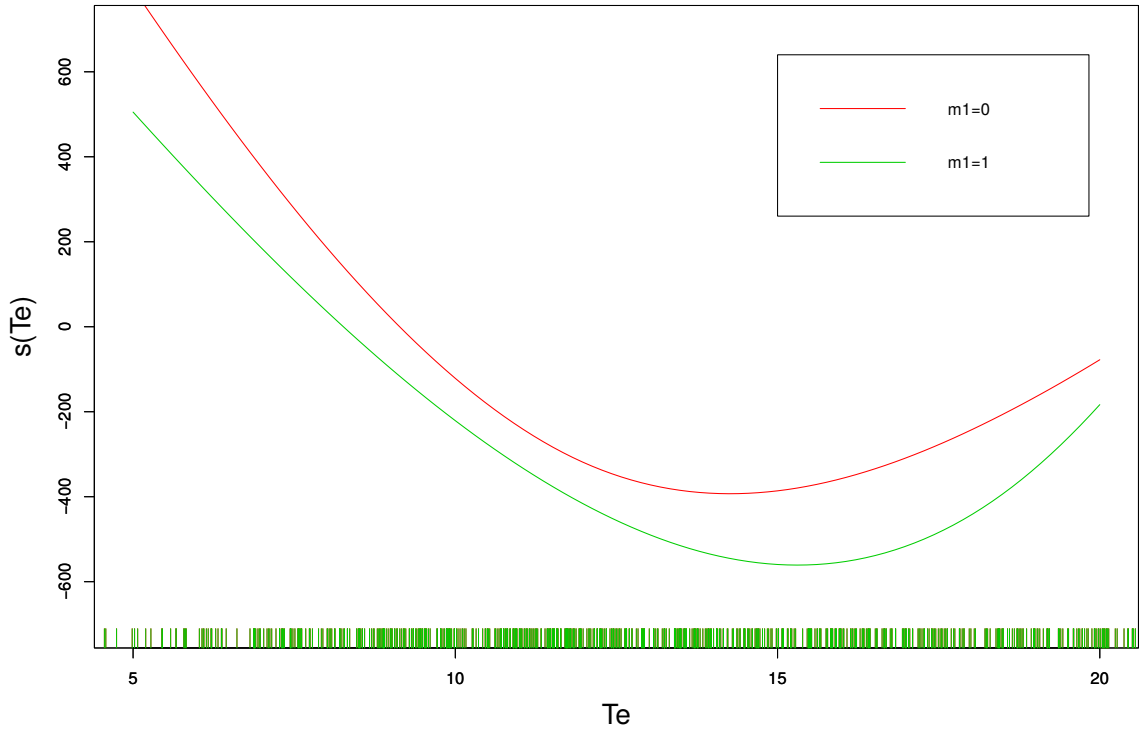

Figure 7.2: Temperature smooth functions before and after the measure

measures are set to 1 only during the heating months). By supposing comparable savings during the two winters we can evaluate yearly $m1$ savings as half of S_{m1}.

By applying equation **??** we can evaluate the impact of the EEM during the heating months:

$$s_{m1} = \frac{S_{m1}}{\bar{E}_y} \cdot \frac{365}{n-k} = \frac{-18648.93}{2756638} \cdot \frac{365}{90} = -0.027 = -2.7\ \% \tag{7.7}$$

Although a better result than the one provided by Model 1.1, where the EEM seemed to be the cause of an increase in the consumption, we can still see that the impact of the measure is quite low. Unfortunately, due to the 'black box' nature of our analysis, it is hard to detect the reasons for this behavior. One important factor to keep into account is that the measure was marked as both a heating and cooling production equipment change. Although, the lack of previous cooling season data without the measure applied, makes it impossible to evaluate the savings coming from the summer use of the equipment, that should be summed to the ones we already calculated, to obtain the total savings connected with this measure.

7.3 Model 2.1

The mathematical equation representing model 2.1 is reported here to allow easier understanding of this section:

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i \omega_i) + \alpha m_1 \tag{7.8}$$

The summary of the model results follows, in terms of R^2 adj., deviance explained and $p - values$:

R^2 adj. $= 0.7$

Deviance explained $= 70.3\%$

Variable	p-value	Variable	p-value
ω_1	$1.87 \cdot 10^{-7}$	ω_2	0.012893
ω_3	0.552468	ω_4	0.173308
ω_5	0.023598	ω_6	0.012190
ω_7	$< 2 \cdot 10^{-16}$	$m1$	0.000921

Table 7.3: P-values Table for Model 2.1

The p-values show high significance for $m1$, by analyzing the value of α_1 and applying equation 6.4 we can evaluate the measure impact on the consumption:

$\alpha_1 = -7.63$

$$s_1 = \frac{\alpha_1}{\bar{E}_d} = \frac{-7.63}{58.54} = -0.13 = -13\ \% \tag{7.9}$$

Having this measure been applied just one month before the last consumption data in our possession, yearly savings can only be calculated as a projection:

$$S_{m1-y} = s_1 \bar{E}_y = (-0.13) \cdot 38440 = -4997.2\ kWh/year \tag{7.10}$$

7.4 Model 3.1

The mathematical equation representing model 3.1 is reported here to allow an easier understanding of this section:

$$E_d = B_l + \sum_{i=1}^{7}(\phi_i \omega_i) + f_1(d_h)_{[m_1=0]} + f_2(d_h)_{[m_1=1]} \tag{7.11}$$

The summary of the model results follows, in terms of R^2 adj., deviance explained and $p - values$:

R^2 adj. $= 0.648$

Deviance explained $= 65.2\%$

Variable	p-value	Variable	p-value
ω_1	$< 2 \cdot 10^{-16}$	ω_2	0.84086
ω_3	0.00587	ω_4	0.88432
ω_5	0.06534	ω_6	$3,75 \cdot 10^{-6}$
ω_7	$< 2 \cdot 10^{-16}$	$f_{m1=0}(d_h)$	0.0841
$f_{m1=1}(d_h)$	0.3712		

Table 7.4: P-values Table for Model 3.1

In this case, not so high values of R^2 adj. and deviance explained, as well as low significance levels for the two splines, proved this model to be unsuitable to evaluate savings for Pilot 3. To perform a more accurate analysis, a K-means clustering algorithm was applied, as explained in section 5.4, its results are presented in the following section.

7.5 Model 3.2

In model 3.2, for the first time a consumption pattern (structure) analysis was realized. First the silhouette method was used to understand the optimal number of structures, which proved to be four.

A K-means clustering analysis performed with $k = 4$ showed that the different structures are not uniformly distributed and that, instead, during the same month, and even during the same week, the building might have different structures. This is probably the reason why the autoregressive variables alone were not able to represent the consumption pattern of the building and yielded such a low R^2 adjusted. In Figure 7.3 we can see the daily structures distribution starting January 2016, while in Figure 7.4 the patterns for the four structures are shown, together with an "NA" category that contains all the daily profiles that could not be associated to any of the 4 clusters.

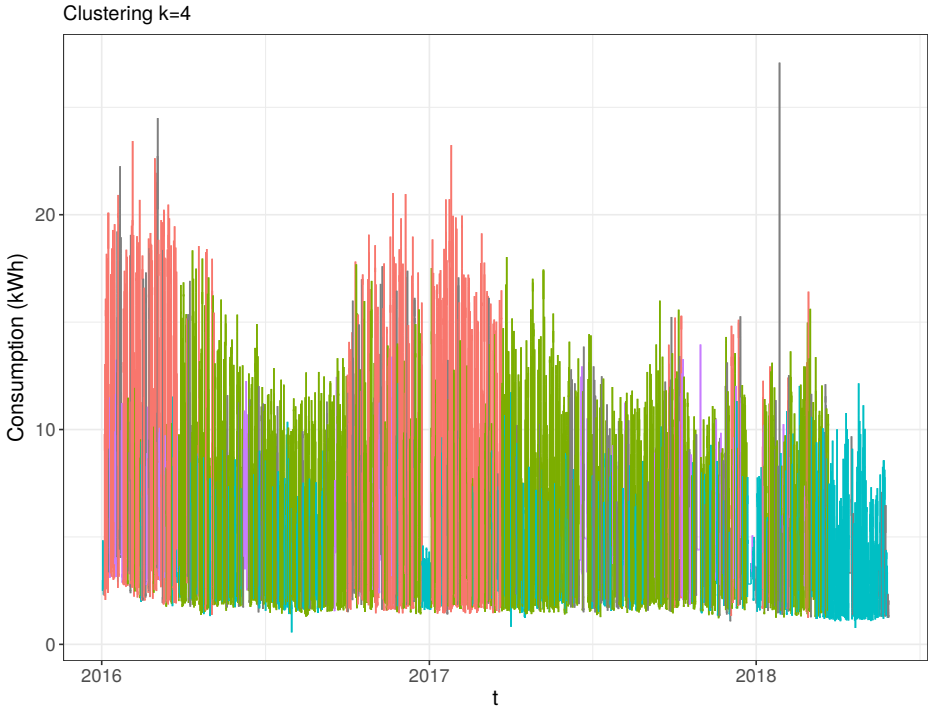

Figure 7.3: Structure distribution over the consumption timeseries

Given the clustering results, the following model was applied:

$$E_d = B_l + \sum_{i=1}^{7}\sum_{j=1}^{k}(\phi_{ij}\omega_{ij}) + \sum_{j=1}^{k} f_j(T_e) + \alpha m_1 \tag{7.12}$$

The model R^2 adj. and deviance explained are presented here, together with the significance level of $m1$ and value of α_1, the complete $p - value$ table is included in the Appendix.

R^2 adj. = 0,836

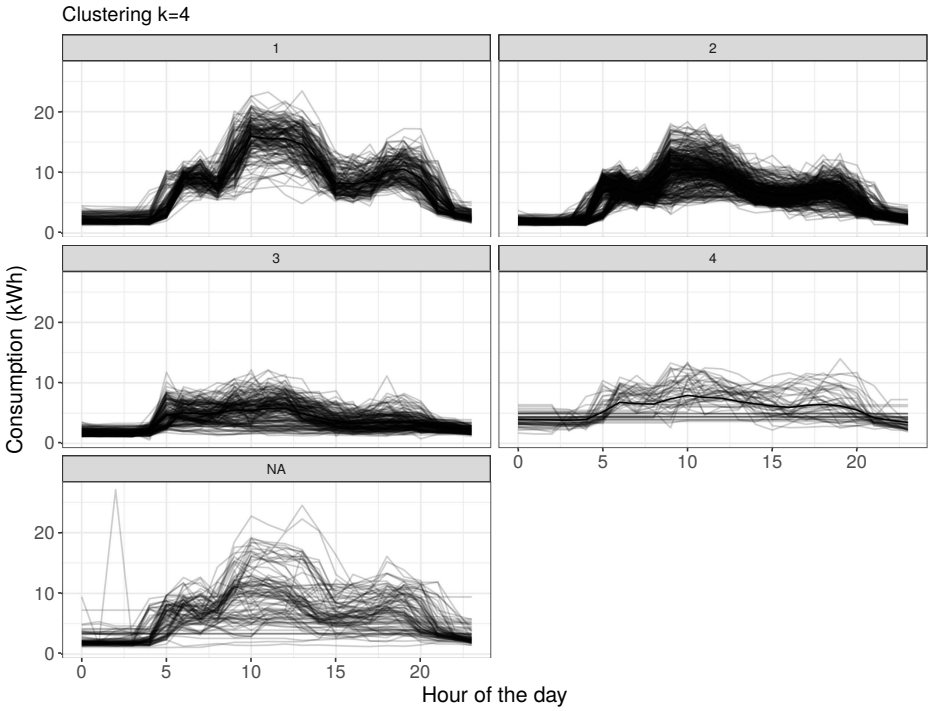

Figure 7.4: Consumption patterns for the four clusters

Deviance explained = 84.3 %

p-value(m1) = $6.49 \cdot 10^{-5}$

$\alpha_1 = -7.42\ kWh$

Applying equation 6.4 we can evaluate the measure impact on the consumption:

$$s_1 = \frac{\alpha_1}{\bar{E}_d} = \frac{-7.42}{139.19} = -0.053 = -5.3\ \% \tag{7.13}$$

Having this measure been applied less than one year before the last consumption data in our possession, yearly savings can only be calculated as a projection:

$$S_{m1-y} = s_1 \bar{E}_y = (-0.053) \cdot 78744 = -4173.4\ kWh/year \tag{7.14}$$

Chapter 8

Conclusions and future work

The research showed that it is possible to evaluate savings using a statistical learning method based on General Additive Models, for tertiary buildings with hourly or sub-hourly consumption data. Although we can not be perfectly sure of the accuracy of our estimations, we can definitely say that the assessed savings are within the range of plausible values, for the considered measures.

The approach proved to be a valid option to easily and cost-effectively assess energy retrofit impact, although the real power of this method lies into the automatization of the process and, in order to implement it, further tests and algorithm developments are needed.

8.1 Future work

The research results represent just the beginning of a long process, in order to allow the method to be applied in a big data environment. This section sums up the further development needed.

In order to validate the model, more buildings need to be analysed with the described method and, when possible, the estimations need to be cross-validated with specific measurements (e.g. monitoring of HVAC units etc.), so to assess the accuracy of the model.

The presented approach works with hourly and sub-hourly data, but as of now, for many buildings we still have access only to monthly consumption data. Evaluating a method to assess energy savings for buildings with monthly data is part of the future work for this thesis. One of the possible solutions would be a linear model that normalizes monthly consumption with heating degree days and cooling degree days. Although, an important aspect is that it is harder to evaluate slight changes in the consumption pattern when only monthly data is available, as we can read in the IPMVP core concepts booklet: "As a rule of thumb, if only monthly billing data are available for energy consumption and demand, savings typically must exceed 10% of the baseline period energy if you expect to confidently discriminate the savings from the unexplained variations in the baseline data."

As we pointed out, the most important step, in order to implement this approach in a big data environment, is to automatize the model application and savings evaluation, that we

manually performed for this thesis. A fundamental part of future work is, therefore, to develop an algorithm able to perform the evaluation with different models, compare them, and choose the one that fits the best the selected building.

Finally, as it was introduced in a previous section, once the model evaluation is automatized and a considerable number of buildings where EEM can be evaluated has been gathered, a neural network for recommendations generation could be implemented.

Bibliography

[1] Luis Pérez-Lombard, José Ortiz, and Christine Pout. A review on buildings energy consumption information. *Energy and Buildings*, 40(3):394–398, January 2008. 1

[2] Pieter de Wilde. The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in Construction*, 41:40–49, May 2014. 7

[3] Ray Galvin. Making the 'rebound effect' more useful for performance evaluation of thermal retrofits of existing homes: Defining the 'energy savings deficit' and the 'energy performance gap'. *Energy and Buildings*, 69:515–524, February 2014. 7

[4] Zhenjun Ma, Paul Cooper, Daniel Daly, and Laia Ledo. Existing building retrofits: Methodology and state-of-the-art. *Energy and Buildings*, 55:889–902, December 2012. 7

[5] Hastie, T. J and Tibshirani, R. J. *Generalized Additive Models*. Chapman & Hall/CRC, 1990. 7

[6] Michael Dibley, Haijiang Li, Yacine Rezgui, and John Miles. An ontology framework for intelligent sensor-based building monitoring. *Automation in Construction*, 28:1–14, December 2012. 8

[7] Kaile Zhou, Chao Fu, and Shanlin Yang. Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews*, 56:215–225, April 2016. 8

[8] Stoyan DANOV, Jordi CIPRIANO, An MEGANCK, Lieven VANDEVELDE, and An Meganck. EMPOWERING CUSTOMER ENGAGEMENT BY INFORMATIVE BILLING – A EUROPEAN APPROACH. page 5, 2015. 12

[9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors. *An introduction to statistical learning: with applications in R*. Number 103 in Springer texts in statistics. Springer, New York, 2013. OCLC: ocn828488009. 17, 18, 19, 20

[10] Sebastian Raschka. *Python machine learning: unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. Community experience distilled. Packt Publishing open source, Birmingham Mumbai, 2016. OCLC: 927507196. 20

[11] C. Fluhrer, E. Maurer, and A. Deshmukh. Achieving radically energy efficient retrofits: The empire state building example. pages 236–243, 2010. 21

[12] Sang Hoon Lee, Tianzhen Hong, Mary Ann Piette, Geof Sawaya, Yixing Chen, and Sarah C. Taylor-Lange. Accelerating the energy retrofit of commercial buildings using a database of energy efficiency performance. *Energy*, 90:738–747, October 2015. 21

[13] Jose L. Molina Felix, Servando Alvarez Dominguez, Laura Romero Rodriguez, Jose M. Salmeron Lissen, Jose Sanchez Ramos, and Francisco J. Sanchez de La Flor. ME3a: Software tool for the identification of energy saving measures in existing buildings: Automated identification of saving measures for buildings using measured energy consumptions. pages 1–6, June 2016. 21

[14] Efficiency Valuation Organization. Core concepts - International Performance Measurement and Verification Protocol. Technical report, 2017. 21

[15] Margaret Fels. PRISM: An introduction. *Energy and Buildings*, 1986. 00307. 21

[16] J. K. Kissock, J. S. Haberl, and D. E. Claridge. Inverse Modeling Toolkit: Numerical Algorithms for Best-Fit Variable-Base Degree Day and Change Point Models. 2003. 00000. 22

[17] Henrik Aalborg Nielsen. *Parametric and non-parametric system modelling*. PhD thesis, Technical University of Denmark, 1999. 00004. 22

[18] Henrik Madsen, Peder Bacher, Geert Bauwens, An-Heleen Deconinck, Glenn Reynders, Staf Roels, Eline Himpe, and Guillaume Lethé. Thermal Performance Characterization using Time Series Data-IEA EBC Annex 58 Guidelines. Technical report, Technical University of Denmark (DTU), 2015. 00003. 23

[19] Hermine Nathalie Akouemo Kengmo Kenfack. *Data cleaning in the energy domain*. PhD thesis, Faculty of the Graduate School, Marquette University, 2015. 00000. 28

[20] José J. López, José A. Aguado, F. Martín, F. Muñoz, A. Rodríguez, and José E. Ruiz. Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electric Power Systems Research*, 81(2):716–724, February 2011. 00026. 41

Appendix

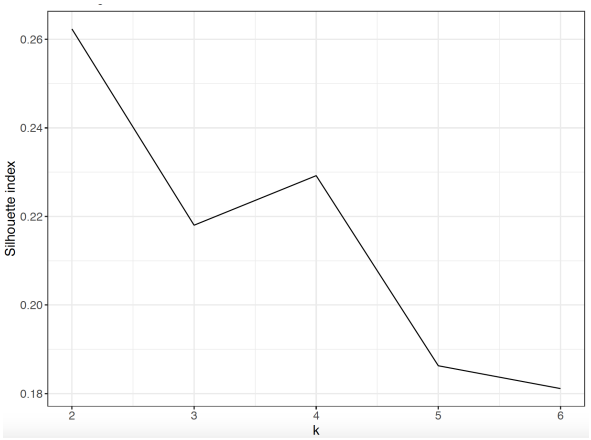

Figure 1: Silhouette index analysis for Belgrave Neighbourhood Centre

Variable	p-value	Variable	p-value
ω_{11}	< 0.054110	ω_{12}	0.000272
ω_{13}	$1.17 \cdot 10^{-6}$	ω_{14}	0.000457
ω_{21}	0.007524	ω_{22}	0.706160
ω_{23}	0.486670	ω_{24}	0.538180
ω_{31}	0.024845	ω_{32}	0.208034
ω_{33}	0.943165	ω_{34}	0.969352
ω_{41}	0.005950	ω_{42}	0.461417
ω_{43}	0.276471	ω_{44}	0.837654
ω_{51}	0.004104	ω_{52}	0.179419
ω_{53}	0.285605	ω_{54}	0.394116
ω_{61}	0.009418	ω_{62}	$5.61 \cdot 10^{-9}$
ω_{63}	0.576789	ω_{64}	0.121814
ω_{71}	0.080129	ω_{72}	0.070566
ω_{73}	0.000102	ω_{74}	0.340092
$f_1(T_e)$	0.000563	$f_2(T_e)$	0.002613
$f_3(T_e)$	0.001626	$f_4(T_e)$	0.730105
$m1$	$6.49 \cdot 10^{-5}$		

Table 1: P-values Table for Model 3.2